DOCUMENT RESUME

ED 385 579 TM 024 017

AUTHOR	Wainer, Howard; And Others
TITLE	How Well Can We Equate Test Forms That Are Constructed by Examinees? Program Statistics Research.
INSTITUTION	Educational Testing Service, Princeton, N.J.
REPORT NO	ETS-RR-91-57; ETS-TR-91-15
PUB DATE	Oct 91
NOTE	29p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Chemistry; Comparative Analysis; Computer Assisted Testing; *Constructed Response; Difficulty Level; *Equated Scores; *Item Response Theory; Models; Selection; Test Format; Testing; *Test Items
IDENTIFIERS	Advanced Placement Examinations (CEEB); *Unidimensionality (Tests)

ABSTRACT

When an examination consists, in whole or in part, of constructed response items, it is a common practice to allow the examinee to choose among a variety of questions. This procedure is usually adopted so that the limited number of items that can be completed in the allotted time does not unfairly affect the examinee. This results in the de facto administration of several different test forms, where the exact structure of any particular form is determined by the examinee. When different forms are administered, a canon of good testing practice requires that those forms be equated to adjust for differences in their difficulty. When the items are chosen by the examinee, traditional equating procedures do not strictly apply. In this paper, how one might equate with an item response theory (IRT) framework is explored. The procedure is illustrated with data from the College Board's Advanced Placement Test in Chemistry taken by a sample of 18,431 examinees. Comparable scores can be produced in the context of choice to the extent that responses may be characterized with a unidimensional IRT model. Seven tables and five figures illustrate the discussion. (Contains 19 references.) (Author/SLD)

RR-91-57

How Well Can We Equate Test Forms That Are Constructed By Examinees?

Howard Wainer
Educational Testing Service

Xiang-Bo Wang
University of Hawai'i

David Thissen
University of North Carolina

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 91-15

Educational Testing Service
Princeton, New Jersey 08541

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

How Well Can We Equate Test Forms That Are Constructed By Examinees?

Howard Wainer
Educational Testing Service

Xiang-Bo Wang
University of Hawai'i

David Thissen
University of North Carolina

Program Statistics Research
Technical Report No. 91-15

Research Report No. 91-57

Educational Testing Service
Princeton, New Jersey 08541

October 1991

How well can we equate test forms that are constructed by examinees?[1]

Howard Wainer
Educational Testing Service

Xiang-Bo Wang
University of Hawai'i

David Thissen
University of North Carolina

Abstract

When an exam consists, in whole or in part, of constructed response items, it is a common practice to allow the examinee to choose among a variety of questions. This procedure is usually adopted so that the limited number of items that can be completed in the allotted time does not unfairly affect the examinee. This results in the de facto *administration of several different test forms, where the exact structure of any particular form is determined by the examinee. When different forms are administered, a canon of good testing practice requires that those forms be equated to adjust for differences in their difficulty. When the items are chosen by the examinee traditional equating procedures do not strictly apply. In this paper we explore how one might equate within an IRT framework. We illustrate our procedure with data from the College Board's Advanced Placement Test in Chemistry.*

[1]This research was supported by the Educational Testing Service through the Research Statistics Project. We are pleased to be able to acknowledge this help. The data, and advice about how to interpret them were supplied by Walter MacDonald, Behroz Maneckshana, Joe Stevens, and Hessie Taft. Without their expert help and kind cooperation we would have had little to write about. We are also grateful to Rick Morgan and Jerry Melican whose careful reviews assured the accuracy of this description. Xiang-Bo Wang's work on this project was done while he was an Educational Testing Service Summer Predoctoral Fellow. We would also like to express our appreciation to R. Darrell Bock whose initial inquires led us to look into this problem.

It has long been understood that a good test must contain enough questions to cover fairly the content domain. In his description of an 1845 survey of the Grammar and Writing Schools of Boston, Horace Mann argued that

> *"... it is clear that the larger the number of questions put to a scholar, the better is the opportunity to test his merits. If but a single question is put, the best scholar in the school may miss it, though he would succeed in answering the next twenty without a blunder; or the poorest scholar may succeed in answering one question, though certain to fail in twenty others. Each question is a partial test, and the greater the number of questions, therefore, the nearer does the test approach to completeness. It is very uncertain which face of a die will turn up at the first throw; but if the dice are thrown all day, there will be a great equality in the number of faces turned up."*

Despite the force of Mann's argument, there are reasons for the increasing pressure to build tests using units that are larger than a single multiple choice item. Sometimes these units can be thought of as aggregations of small items, e.g., testlets (Wainer & Kiely, 1987; Wainer & Lewis, 1990); sometimes they are just large items (e.g., essays, mathematical proofs, etc.). Large items, by definition, take the examinee longer to complete than do short items. Therefore, fewer large items can be completed within a given testing time. The fact that an examinee cannot complete very many large items within the allotted testing time places the test builder in something of a quandary. One must either be satisfied with fewer items, and possibily not span the content domain as fully as might have been the case with a much larger number of smaller items, or expand the testing time sufficiently to allow the content domain to be well represented. Often practicality limits testing time, and so compromises on domain coverage must be made. A common compromise is to provide several large items and allow the examinee to choose among them. The notion is that in this way the examinee is not placed at a disadvantage by an unfortunate choice of domain coverage by the test builder.

Allowing examinees to choose the items they will answer presents a difficult set of problems. Despite the most strenuous efforts to write items of equivalent difficulty, some might be more difficult than others. If examinees who choose different items are to be fairly compared with one another, the scores obtained on those items must be equated. How?

All methods of equating are aimed at producing the subjunctive score that an examinee would have obtained had that examinee answered a different set of items. To accomplish this feat requires that the item responses that are not observed are "missing-at-random." The act of equating means that we believe that the performance that we observe on one item tells us something about what performance would have been on another item. If we know that the procedure by which an item was chosen has nothing to do with any specialized knowledge that the student possesses we can believe that the missing responses are missing-at-random. However, if the examinee has a hand in choosing the items this assumption becomes considerably less plausible.

To understand this more concretely consider two different construction rules for a spelling test. Suppose we have a corpus of 100,000 words of varying difficulty, and we wish to manufacture a 100-item spelling test. From the proportion of the test's items that the examinee correctly spells we will infer that the examinee can spell a proportion of the total corpus. Two rules for constructing such a test might be:

- *Missing-at random*: We select 100 words at random from the corpus and present them to the examinee. In this instance we believe that what we observe is a reasonable representation of what we did not observe.

- *Examinee selected*: A word is presented at random to the examinee, who then decides whether or not to attempt to spell it. After 100 attempts the proportion spelled correctly is the examinee's raw score. The usefulness of this score depends crucially on the extent to which we believe that examinees' judgments of whether or not they can spell particular words are related to actual ability. If there is no relation between spelling ability and *a priori* expectation, then this method is as good as missing-at-random. At the other extreme, we might believe that examinees know perfectly well whether or not they can spell a particular word correctly. In this instance a raw score of 100% has quite a different meaning. Thus, if an examinee spells 90 words correctly all we can be sure of is that that examinee can spell no fewer than 90 words and no more than 99,990. A clue that helps us understand how to position our estimate between these two extremes is the number of words passed over during the course of obtaining the sample of 100. If the examinee has the option of omitting a word, but in fact attempts the first 100 words presented, our estimate of that examinee's proficiency will not be very different than that obtained under 'missing-at-random.' If it takes 50,000 words for the examinee to find 100 to attempt we will reach quite a different conclusion. If we have the option of forcing the examinee to spell some previously rejected words (sampling from the unselected population), we can further reduce uncertainty due to selection.

This example should make clear that the mechanism by which items are chosen is almost as crucial for correct interpretation as the examinee's performance on those items. Is there any way around this problem? How can we equate tests in which all, or some, of the items are selected by the examinee? In this paper we examine one strategy for equating different test forms that are constructed through examinee choice.

The data

The data we use to illustrate our methodology are from the 1989 Advanced Placement Examination in Chemistry. The Advanced Placement (AP) Program of the College Board is meant to evaluate the efficacy of college level courses taught in secondary schools. Exams are given in a variety of subjects and validity studies have been done that have established a relationship between performance on these tests and likely performance in associated college courses. A general finding is that AP students generally do better in advanced college "courses than do the students who have taken the regular freshman-level courses at that institution." (p. v, *The 1989 Advanced Placement Examination in Chemistry and its grading*).

The 1989 Advanced Placement Examination in Chemistry is three hours long. It is divided into two sections with 90 minutes allotted for each. Section I consists of 75 five-option multiple choice questions and accounts for 45% of the total grade. Section II consists of problems and essay questions, and has four parts:

Part A is a single problem (Problem 1) that all examinees must answer, and accounts for 14% of the total grade.

Part B has two problems (Problems 2 and 3), and the examinee must answer exactly one of those. This part accounts for 14% of the total grade.

Part C is treated as a single problem (Problem 4), but has eight parts. The examinee must choose five of these to answer. This part accounts for 8% of the total grade.

Part D has five problems (Problems 5, 6, 7, 8 and 9) out of which the examinee must answer three. This part accounts for 19% of the total grade.

This form of the exam was taken by approximately 18 thousand students[2] in 1989. The test form has been released and interested readers may obtain copies of it with the answers and a full description of the scoring methodology from the College Board.

One can think of the various choice options yielding different test forms. Thus a student who opted to answer Problem 2 had a different form of the test than a student who opted for Problem 3. These two students would have in common the 75 multiple choice items as well as Problem 1. The exam, as currently configured, can be partitioned by examinee choice into 1,120 forms[3] that overlap.

---

[2]The actual number of examinees was 18,462; however 31 tests were handed-in essentially blank and so were excluded from the analysis.

[3]Part B yields $\binom{2}{1} = 2$ possible choices, part C yields $\binom{8}{5} = 56$ choices, and part D yields $\binom{5}{3} = 10$ choices. The product of these yields the total possible number of different test forms.

At this point it is important to emphasize the crucial difference between the problem of equating these 1,120 examinee-selected forms versus what could be done if the different forms were assigned at random to examinees. In the latter case any one of a number of traditional common-item equating methods would work. In this instance the violation of the assumption that unseen responses are 'missing-at-random' can make those methods completely invalid. To equate we must make some assumptions about the missing data. Unfortunately, there is nothing in the observed data that can allow us to test the validity of such assumptions.

Consider a simplified, but extreme case. Suppose we truncate the test after Problem 3. We now have two groups of examinees, those who answered Problem 2 (Group I) and those who answered Problem 3 (Group II). Both groups were presented with the 75 multiple choice items and Problem 1. If we utilize traditional equating methods we could obtain an estimate of the score that those in Group I would have obtained on Problem 3. Suppose further that chemistry can be taught in two, quite different ways[4] and that the choice of questions reflects this diversity of course content. Examinees would tend to choose the question that reflected their course of instruction. The subjunctive inference made from the equating ignores the possibility that examinees might have a pretty good idea of what kinds of questions they would have a better chance on—that their nonresponse is nonignorable in the sense of Rubin (1987). Thus, when we speak of equating in this instance we mean that the missing score that is estimated by the equating model is the score that the examinee would have obtained had that examinee taken the course for which that item choice was better suited. If you think that such inferences are a bit of a stretch, you are beginning to understand the logical dilemma of equating test forms involving examinee choice.

The data analysis plan

The goal of this investigation is to examine the effects of examinee choice on scores on the Chemistry AP test. We do not believe that it is possible to obtain unequivocal answers because the nature of nonignorable nonresponse implies that available information is insufficient. One must either gather additional information (perhaps validity data) or assume away the problems. This study examines the viability of several plausible assumptions.

We will consider the following questions:

1. Are we measuring the same thing with essays/problems as we are with multiple choice questions?

2. Is the current method of scoring examinee chosen items viable (are the choices of equal difficulty)?

3. If they appear not to be of equal difficulty how can we equate them?

---

[4]One approach might emphasize all chemical reactions in terms of energy considerations, a second might be the traditional approach involving valences and equation balancing.

We will examine each of the questions in turn. However, before presenting all of the details let us describe the analysis plan in broad terms. First, we will confine ourselves to the study of a truncated test consisting only of the 75 multiple choice items and Problems 1, 2 and 3. A generalization of our methodology to the entire 1,120 possible forms is theoretically straightforward, although practically daunting.

For question 1:

We will fit an IRT model to Problems 1, 2 and 3 jointly, assuming ignorable nonresponse for the moment—Problem 1, in this instance, provides the link for common-item equating. This result, given the assumptions, yields an estimate of test forms equated on an "essay proficiency."

Second, we redo this analysis, but use a set of multiple choice items as the link between Problems 2 and 3. This yields an estimate of test forms equated on a "multiple choice proficiency."[5]

If the parameters of Problems 2 and 3 are the same regardless of which anchor is used we can conclude that there is no evidence to believe that essay proficiency is different from multiple choice proficiency.

For question 2:

We examine the fit (-2loglikelihood) of a model that allows both Problems 2 and 3 to have different parameters and compare it with a model that restricts the two Problems' parameters to be equal. If there is a significant increase in the quality of the fit with the more general model we will conclude that the current scoring scheme that treats the problems interchangeably is not justified. We will examine the size of the inequity by calculating the expected raw score that each examinee would have on each of the two items. This methodology is formally identical with that developed by Wainer, Sireci & Thissen (1991) for detecting and measuring testlet DIF.

For question 3:

The very essence of nonignorable nonresponse is that, if it exists, one cannot examine its effects with the data in hand. The answer to this question involves estimating how well an examinee would have done on a problem that they opted not to answer. We simply have no statistical way to assess how well any psychometric model does this.

[5]Actually it is not pure 'multiple choice' since the estimates of θ are a mixture of multiple choice items and a single essay, but if we are testing different things this should yield a different result than was obtained with Problem 1 as the anchor.

Instead, we will use the multiple choice anchor and the performance of Group I on Problem 1 to predict how well someone in Group II would do on Problem 1. We can do this by temporarily deleting the scores of individuals in Group II on Problem 1, repeating some of the analyses performed to answer questions 1 and 2, and comparing the estimated performance of Group II with their actual performance. We can also do the same thing for Group I. While this does not tell us how well our model for equating works on problems for which we do not have a criterion, it does provide a plausible upper bound on its accuracy.

The details

The statistical model we used is an IRT model developed by Bock (1972) for fitting data in nominal categories. It specializes to a model for ordered categories by imposing monotonicity constraints on some of its parameters. We have used this successfully in a variety of contexts (Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg & Fitzpatrick, 1989; Thissen, Steinberg & Mooney, 1989; Thissen, Steinberg & Wainer, 1992; Wainer, Sireci, & Thissen, 1991).

Bock's 1972 Model

Suppose we have J large items, indexed by j, where $j = 1, 2, ..., J$. On each item there are m_j possible scores, so that for the jth item there is the possibility of responses $x_j = 0, 1, 2, ..., m_j$. The statistical scoring model posits a single underlying (and unobserved) dimension that we call latent proficiency, and denote θ. The model then represents the probability of obtaining any particular score as a function of proficiency. For each item there is a set of functions, one for each response category. These functions are sometimes called item characteristic curves (Lord & Novick, 1968), item operating curves (Samejima, 1969), or trace lines (Thissen, Steinberg & Mooney, 1989). We shall follow Thissen et al.'s (1989) notation and nomenclature.

The trace line for score $x = 0, 1, ..., m_j$, for item j is

$$T_{jx}(\theta) = \frac{\exp[a_{jx}\theta + c_{jx}]}{\sum_{k=0}^{m} \exp[a_{jk}\theta + c_{jk}]} ; \qquad (1)$$

where the $\{a_k, c_k\}_j$, $k = 0, 1, ..., m_j$ are the item category parameters that characterize the shape of the individual response trace lines. The a_ks are analogous to discriminations; the c_ks analogous to intercepts. The model is not fully identified, and so we need to impose some additional constraints. It is convenient to insist that the sum of each of the sets of parameters equal zero, i.e.

$$\sum_{k=0}^{m_j} a_{jk} = \sum_{k=0}^{m_j} c_{jk} = 0 \quad .$$

In this context, we reparameterize the model using centered polynomials of the associated scores to represent the category-to-category change in the a_ks and the c_ks:

$$a_{jk} = \sum_{p=1}^{P} \alpha_{jp} \left(k - \frac{m_j}{2} \right)^p \tag{2}$$

and

$$c_{jk} = \sum_{p=1}^{P} \gamma_{jp} \left(k - \frac{m_j}{2} \right)^p \tag{3}$$

where the parameters $\{\alpha_p, \gamma_p\}_j$, $p = 1, 2, \ldots P$, for $P \leq m_j$ are the free parameters to be estimated from the data. The polynomial representation has, in the past, saved degrees of freedom with no significant loss of accuracy. It also provides a check on the fit of the model when the categories are ordered. If the categories are ordered the a's must be monotonically ordered (Wainer, Sireci & Thissen, 1991).

The analysis strategy

We initially treat the multiple choice and the 'essay' parts of the exam separately. However, after the initial analyses they will be joined as needed. Much of our argument about the structure of the 'missingness' of responses will be based on plausible but untestable assumptions. We will try to be explicit about these through the ensuing discussion.

Statistical decisions, when done on the basis of formal hypothesis tests, are based on the size of the likelihood ratio. We fit two models, one a proper submodel of the other, and compare their likelihoods (actually we will look at the difference between the chi-square statistic obtained from -2loglikelihood for each model). The difference between two chi-squares is also a chi-square whose degrees of freedom is the number fewer parameters in the more restricted model. This methodology is widely used; see Judd & McCleliand (1989), who base an entire statistics course on this concept.

We shall not be overly concerned with overall statistical fit measures because a sample of more than 18 thousand examinees allows us to reject most models. Instead, we concentrate primarily on the absolute size of effects and the size of differences.

The size of the test is unwieldy for the large number of analyses required in developmental work like this. Consequently we will follow Einstein's advice[6] and simplify matters considerably. In doing so we will try to maintain the minimal test that still contains the elements that we wish to investigate. The first simplification is to reduce the essay portion of the test to parts A and B. This partitions the examinee population into two self-selected groups (Groups I and II, who chose respectively, Problems 2 or 3). We opted for this for several reasons. First because it carried the selection aspect of the problem. Second, it allowed the development of a methodology that could easily be expanded to include a greater number of examinee partitions.

There are some complications with the data set that could not be removed, but whose effects are limited. One of these problems involves zero scores. The most common score on these problems was zero. Sometimes this is because the examinee tried the problem but failed to achieve any creditable result. Sometimes it is because the examinee omitted the problem entirely. When this occurs we have no information about which problem (2 or 3) was omitted, and the data entry process assigned a zero score, more-or-less randomly to one or the other choice. A large part of the small misfit of the model is due to an overabundance of zero scores.[7]

While we examined the entire multiple choice portion of the exam, we have opted to use only six items from this exam for the multiple choice anchor. In the past (Wainer et al, 1991) we have found that three or four well chosen items are sufficient for this purpose. Thus we are assured that six would be ample. The analyses that we report below confirm this.

The Results

The multiple choice section

The first analysis calibrated the entire multiple choice exam. We fit its 75 items with a three parameter logistic IRT model using the computer program BILOG (Mislevy & Bock, 1983). The section's marginal reliability was over .91; the mean difficulty was .49, the mean slope was .73, and the mean lower asymptote was .17. In general it looked just like many other high quality, professionally produced, exams.

We examined the dimensionality of the items using full information factor analysis (Bock, Gibbons, & Muraki, 1988) and discovered that, although it required three dimensions to obtain an acceptable fit, those three dimensions were highly intercorrelated and a one-dimensional solution did very well indeed. Trimming off some of the late-appearing items that were not reached by a significant proportion of the sample further strengthened the one dimensional solution.

---

[6]"Everything should be as simple as possible, but no simpler."

[7]Current practice corrects this problem with graders assigning a '-' for an omission and hence reserving a score of '0' for 'attempted, but failed to achieve any credible result.'

From our analyses of the multiple choice section we chose six representative items to serve as an anchor in some of the subsequent analyses. These items were selected on the basis of six criteria:

1. Their difficulties spanned the plausible range of the examinees' proficiency distribution.
2. They were as discriminating as possible.
3. Their lower asymptotes were low enough to assure that guessing was minimal.
4. They fit the IRT model well and were situated as near to the principal dimension obtained from the factor analysis as possible.
5. They were attempted by as large a proportion of the examinee population as possible.
6. They, *in toto,* should be reasonably representative of the section's content domain.

It was not possible to find items that satisfied all of these goals simultaneously, because it is hard to write items at extreme difficulties with steep slopes. Moreover, difficult items tend to be at the end of the section and are often compromised by increased omission. There were four content areas represented on the test; we managed to include a representative of three of these in our sample. The six items chosen as well as their statistical and content characterizations are shown in Table 1.

Table 1. *Characteristics of the six items chosen to serve as the equating anchor*

Item Number	Slope	Difficulty	Lower Asymptote	Percent Reached	Item Content
1	0.85	-1.08	0.16	100	Descriptive chemistry & lab
40	1.07	-0.62	0.15	100	Reaction
13	1.04	0.61	0.16	100	Structure of matter
55	1.02	1.13	0.22	100	Reaction
66	1.13	1.70	0.20	96	Reaction
69	0.72	3.34	0.21	94	Descriptive chemistry & lab

Why did we restrict ourselves to only six items for an anchor? Why not use all, or most, of the 75? There are two parts to the answer to these questions. First, we expect that for our purposes six items will provide ample stability for an anchor test. This has been the case in earlier work on the development of a 'designated anchor' in DIF work (Wainer et al, 1991). Secondly, in subsequent analyses we will be using the response pattern on the

anchor, not the raw score. Six items add 2^6 (= 64) patterns to the analysis. When we try to equate two problems, each having 10 score categories with an anchor made up of n binary items, the table analyzed is $2 \times 10 \times 2^n$. For 6 items this is 1,280, for 7 items 2,560, and for 8 items 5,120. It is clear that even with 18 thousand examinees one very quickly reaches a point at which there are many cells with few observations. Thus unless we are willing to use 'number right' as the stratifying variable we face serious practical constraints. 'Number right' is justified (that is, a sufficient statistic for the estimation of θ) only for the Rasch model. Because of the frequency of guessing, these items are not well fit with the Rasch model.

Constructed Response Section
Problems 1, 2 and 3

The first three problems (Sections A and B) were fit with the IRT model in equation 1, using the computer program Multilog (Thissen, 1991). This was done in three stages. First we fit those 14,270 examinees who answered Problems 1 and 2 (Group I) and then, separately, the remaining 4,161 who answered Problems 1 and 3. These separate analyses provided us with independent estimates of the parameters for Problem 1 and fit statistics for the two groups separately. We then specialized the model for each by reducing the power of the polynomials for the '*a*' and '*c*' parameters for each problem. During the course of this we discovered that with this large sample size, small perturbations in the fit of the polynomial to the estimated parameters can have a profound effect on the likelihood. This was different from our previous experience in other contexts. Consequently we sacrificed the possible savings of some degrees of freedom and over-fit the model. For example, the polynomial used to fit the the ten values of '*c*' for each model was of 9th degree. The *a*'s were fitted with quartics.

After doing this initial calibration we joined the two Groups and redid the analysis with both groups together. The form of the data is shown in Table 2.

Table 2. *Schematic structure for equating problems 2 and 3*

Problem	Group I	Group II
1	X	X
2	X	
3		X

From this analysis we obtained an estimate of the difference between the means of the proficiency distribution of the two groups as well as equated estimates of the parameter values for all three items. These estimates are on the same IRT scale, given Problem 1 as the anchor. Crucial to the interpretation of these results is the untestable assumption that conditioning on θ through Problem 1 accounts for any differences observed in the groups due to their choice behavior. We will examine further the viability of this assumption shortly. The results of this analysis are shown in Table 3.

Table 3. Estimated parameters for Problems 1, 2, and 3

Score Category	*Problem 1*		*Problem 2*		*Problem 3*	
	a	*c*	*a*	*c*	*a*	*c*
0	-4.5	1.2	-4.8	0.7	-3.1	0.5
1	-2.8	2.0	-3.6	1.1	-2.2	1.3
2	-1.5	2.5	-2.5	1.4	-1.4	1.2
3	-0.6	2.6	-1.6	1.5	-0.7	1.1
4	0.2	1.6	-0.6	1.3	-0.2	0.9
5	0.8	0.2	0.4	0.9	0.3	0.5
6	1.4	-1.1	1.5	0.3	0.8	0.1
7	1.9	-2.2	2.7	-1.1	1.4	-0.8
8	2.3	-2.9	3.8	-2.4	2.0	-1.6
9	2.7	-3.8	4.8	-3.8	3.0	-3.3

The mean of Group II was fixed at 0 and the mean of Group I was estimated to be -.03. This small estimated difference in mean θ suggests that the two groups formed on the basis of their choice in Part B of the exam are of roughly the same proficiency on the IRT scale.

At this point we can answer several interesting questions from the values of the fit statistics in the various analyses. Shown in Table 4 are the values of G^2 obtained from -2loglikelihood. While their absolute values are hard to interpret, comparisons among them are meaningful. To begin, we note that when we add the G^2 associated with the fit of Group I alone to that obtained from Group II alone we obtain a value of 196. When we fit data from the two groups together we are estimating only one set of *a* and *c* parameters for Problem 1 instead of the two sets used when they were analyzed separately. The difference between the fit with this equality constraint and without it, 216 – 196 = 20, is distributed as a χ^2 on 12 degrees of freedom and is not significant. This implies that Problem 1 performs the same in both groups (no DIF).

Table 4. Summary of fit statistics for the sequence of models tested

Model	*d.f.*	G^2
Group I Alone	73	118
Group II Alone	73	78
Total	146	196
Groups I & II Jointly:		
2 & 3 parameters estimated	158	216
A2=A3 constrained	162	311
A2=A3 & C2=C3 constrained	171	508

The next three analyses were meant to determine the viability of the current practice of treating Problems 2 and 3 as equally difficult. If this practice is valid we should find that when we constrain the parameters of Problems 2 and 3 to be equal the value of the goodness-of-fit G^2 should not increase much more than the number of degrees of freedom. The first analysis constrained all of the *a's* and *c's* for Problems 2 and 3 to be equal. This resulted in a G^2 of 508, an increase of 292. This leaves no doubt that these two problems are not performing the same. The next analysis constrained the *a*'s separately in an effort to determine what characteristic of the problems differed (slope or intercept). As is evident from the results shown in Table 4, the answer is "both."

These analyses speak to the statistical significance of the observed differences between the item parameters for Problems 2 and 3. The next question we must address is the size of these differences. The parameters shown in Table 3, when substituted into Equation (1), generate a set of 10 trace lines for each item. Each trace line is the probability of obtaining a particular score as a function of proficiency (θ). We can easily aggregate across trace lines by multiplying the value of the trace line by the score category and summing across trace lines to obtain the expected score as a function of θ. That is,

$$E(\text{Score}|\theta) = \sum_{j=0}^{m} x T_{jx}(\theta) \quad . \tag{4}$$

Shown in Figure 1 are the expected score curves for Problems 2 and 3. As is evident Problem 2 is easier than Problem 3 for examinees with above average proficiency. The exact amount of the advantage for examinees who chose Problem 2 is easily seen when we plot the difference between the two curves. Figure 2 shows that there is about a one point advantage to choosing Problem 2 for an examinee whose proficiency is about θ=1.5.

Figure 1

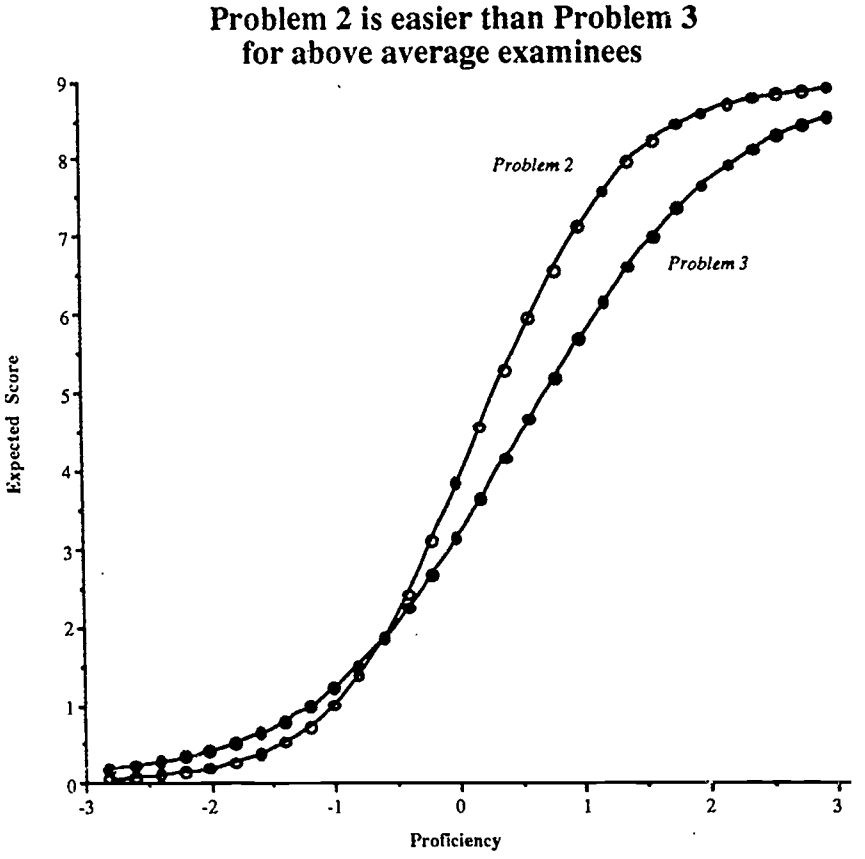
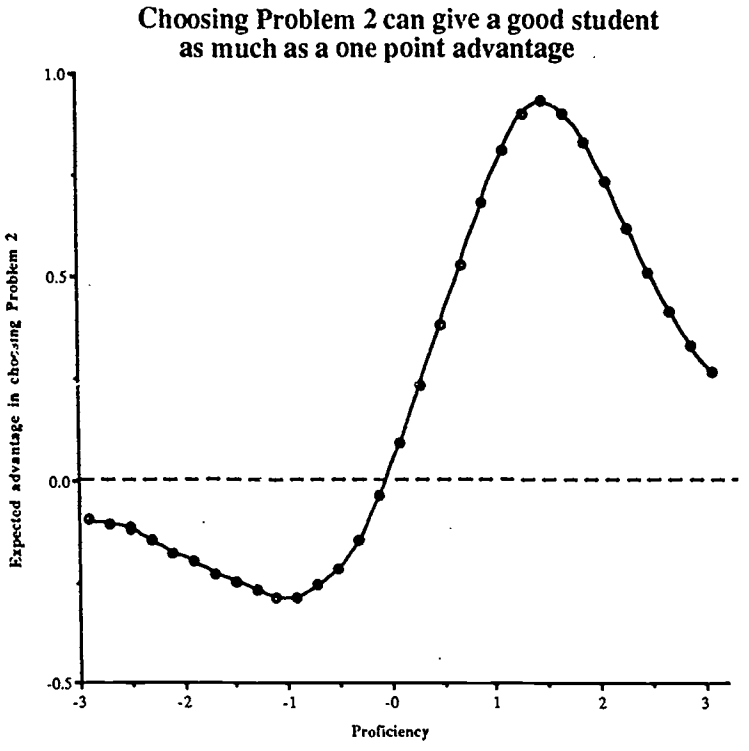

Problem 2 is easier than Problem 3
for above average examinees
Problem 2
Problem 3
Expected Score
Proficiency
0
1
2
3
4
5
6
7
8
9
-3
-2
-1
0
1
2
3

Figure 2



Combining Multiple Choice and Constructed Responses

The next step in our investigation involves combining the information contained in the multiple choice section with that in the constructed response section. We do this by using the six multiple choice items as the linking items to calibrate Problems 2 and 3. This is identical to the procedure described in the previous section except that the 6 multiple choice items are substituted for Problem 1. We do this for several reasons. First to determine the extent to which the trait being tested by the multiple choice items is the same as that being tested by the constructed response items. If it is the same trait then certainly many of the measurement goals of the test would be satisfied more economically by expanding the multiple choice section. More to the point of this investigation, if the multiple choice anchor works it eases many practical problems associated with equating all of the possible 'forms' of this test.

We have found that one way to think about these results is that there is a 'constructed response θ' and a 'multiple choice θ.' We wish to know how similar these two latent dimensions are. We determine this by considering the response curves for Problems 2 and 3 when they are calibrated on these two latent dimensions separately. The

extent to which they are similar characterizes the extent to which the two possible latent dimensions are really the same. We found that the two θs appear to be closely related. To the extent that any difference exists it involves Problem 2. As shown in Figure 3, the relatively steeper slope of the expected score curve for Problem 2 calibrated with Problem 1, as opposed to the same curve when it is calibrated with the multiple choice items, indicates that Problem 2 is somewhat more closely related to Problem 1 than it is to the multiple choice items. This is exactly what one would expect if the 'multiple choice θ' was not quite the same as the 'constructed response θ.' Remember that this curve would be horizontal if they were orthogonal; that is if there was no relation between multiple choice proficiency and expected score on a constructed response problem.

Figure 3

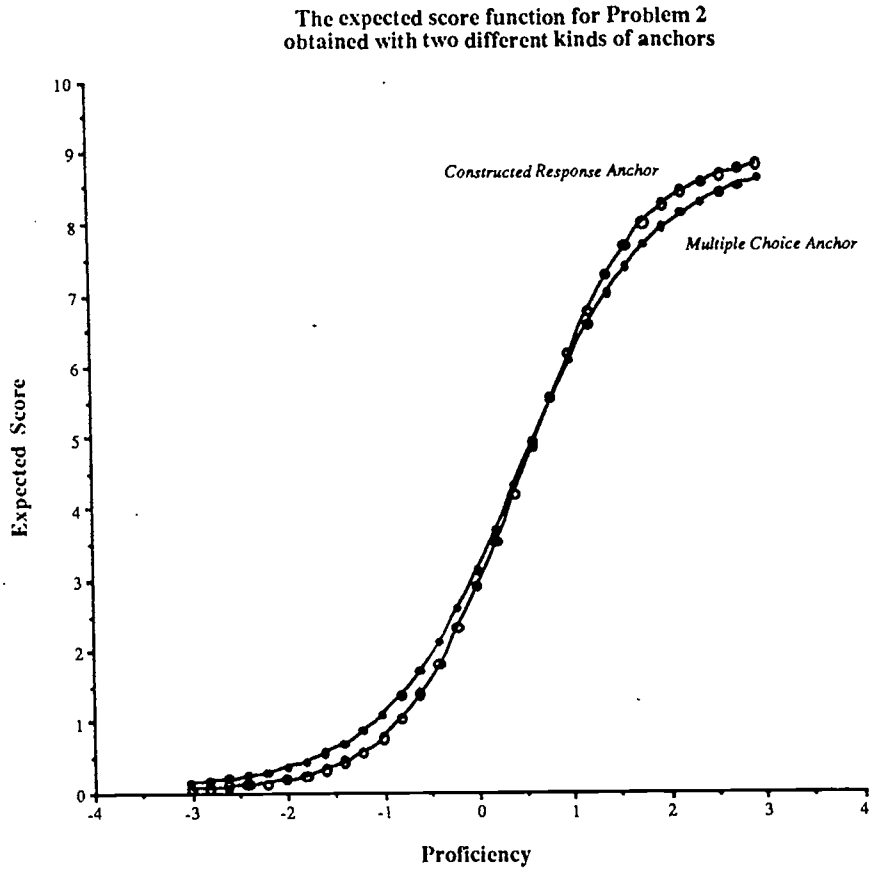
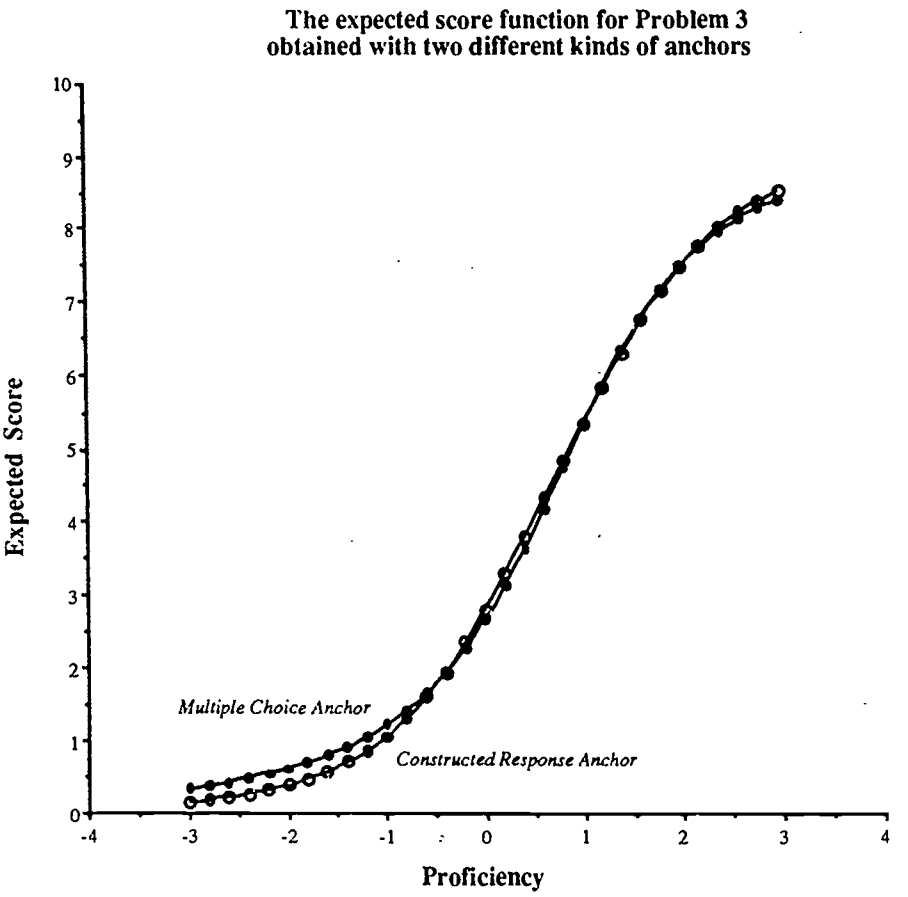


Figure 4



The expected score curves for Problem 3, shown in Figure 4, are nearly identical. We conclude that we will not go very far wrong acting as if both kinds of items are testing essentially the same thing. Certainly a comparison of the difference between the expected curve lines for Problems 2 and 3 (shown in Figure 1) is very large in comparison to the differences seen in Figures 3 and 4. It is far less cavalier to believe that multiple choice items test the same thing as do constructed response items than to believe that Problems 2 and 3 are equally difficult.

What would IRT scale scores be?

IRT scale scores directly correct for the differences in difficulty between the two Problems. The *expected a posteriori* (EAP) estimates of θ for several illustrative response patterns are shown in Table 5. For response patterns associated with very low levels of θ,

like 00 on Problems 1-2 or 1-3, or 000000 on the six multiple choice items and 0 on the constructed response problems, there is very little difference between the estimates involving Problems 2 and 3. However, for response patterns associated with higher levels of θ, the IRT scale score is about 0.15 standard units higher for the reponse pattern involving the more difficult Problem 3 than it is for the corresponding response pattern involving Problem 2.

***Table 5**. An illustration of the value of q obtained for various response patterns*

Response Pattern	EAP[θ] Problems 1-2	EAP[θ] Problems 1-3	Response Pattern	EAP[θ] M. Choice Problem 2	EAP[θ] M. Choice Problem 3
00	–1.39	–1.41	000000 0	–1.59	–1.57
33	0.15	0.30	111000 4	0.15	0.22
77	1.35	1.52	111110 7	1.26	1.37
99	2.17	2.29	111111 9	2.17	2.31

What is the effect of examinee choice?

So far we have been assuming that the examinees' choice is, in some sense, at random. Empirically what this assumption means is that we could:

(i) Estimate an examinee's proficiency from those parts of the test that were chosen,

(ii) Estimate the parameters of each item from those examinees who chose it , and

(iii) Use the IRT model to describe the performance of each examinee on those items that were not chosen.

Goodness-of-fit and other statistical measures do not directly speak to the accuracy of predictions about what has not been observed. Strictly speaking we cannot go any further without gathering data from examinees on the items that they chose not to answer and comparing these results with what was predicted from the model. This is at least impractical and may be impossible.

However, we can do something with the data on hand that provides a plausible upper bound on the quality of the equating scheme. We can omit part of the data for one of the two groups and predict what those scores would have been, and then see what they in fact were. Thus we might set aside Group II's data on Problem 1 and estimate performance

on that problem based on the multiple choice anchor and the performance of Group I on Problem 1. Of course we could do the same thing with the groups reversed. One obvious approach is to compare the parameter structure for Problem 1 obtained with each group separately, since it is those parameters that determine the response pattern frequencies. How can we estimate the parameters for Problem 1 separately for Groups I and II using only the multiple choice items as anchor? And, more specifically, how can we determine if the two different estimates of those item parameters are significantly different from one another?

We used established DIF technology (Thissen, Steinberg & Wainer, 1992; Wainer Sireci & Thissen, 1991) to answer the question, "Does Problem 1 operate differently in Group I than in Group II?" The structure of the analysis is summarized in Table 6, in which we denote Problem 1 for Group I as "1*" and for Group II as "1**."

Table 6. *Schematic represntation of the analysis plan for examining the credability of the ignorable nonresponse assumption with problem 1, using the multiple choice items as the equating anchor.*

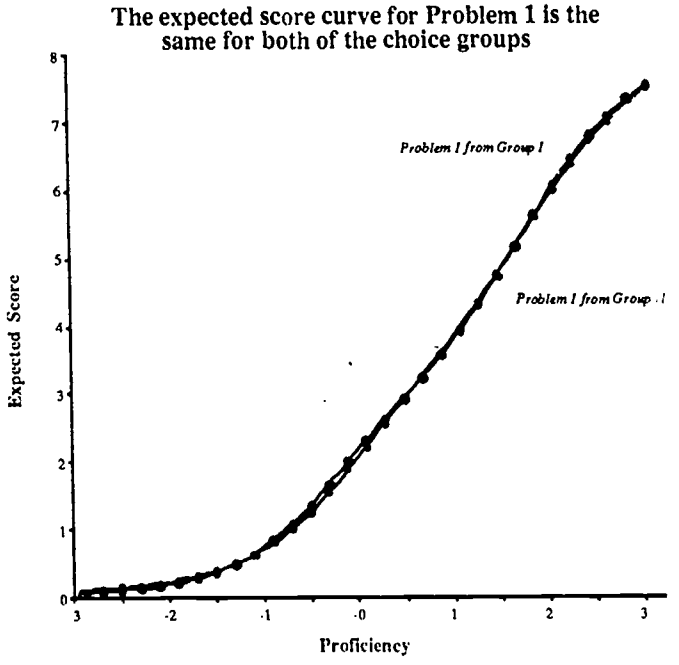
Problem	Group I	Group II
MC	X	X
1*	X	
1**		X

In the analysis, we act as though examinees in Group I answered Problem 1* but omitted Problem 1**, and that those in Group II did the opposite. The 6 multiple choice items act as the anchor and the parameters for Problem 1 are estimated separately for the two groups. The estimated value of the parameters obtained as well as the likelihood ratio G^2 are shown in Table 7. The rightmost panels of Table 7 contain the estimated parameters for Problem 1 when they are constrained to be equal in the two groups (the "No DIF" model). As is evident from the values of the parameters, there is very little difference in the performance of Groups I and II on Problem 1. The likelihood increases by 22 (on 13 degrees of freedom), which confirms this impression. Last, in Figure 5, we show the plots of expected scores obtained from the model for the parameter estimates obtained separately for each group: the two curves are virtually coincident.

Table 7. The estimated parameters obtained for Problem 1 under three different circumstances.

Score Category	***Problem 1****		***Problem 1****		***Problem 1***	
	a	*c*	*a*	*c*	*a*	*c*
0	-3.9	1.5	-3.6	1.7	-3.8	1.6
1	-2.7	2.0	-2.7	2.0	-2.7	2.0
2	-1.7	2.5	-1.7	2.3	-1.7	2.5
3	-0.8	2.6	-0.7	2.4	-0.8	2.6
4	0.0	1.6	0.1	1.5	0.1	1.6
5	0.8	0.2	0.8	0.0	0.8	0.1
6	1.4	-1.3	1.3	-1.1	1.4	-1.2
7	2.0	-2.4	1.8	-2.2	1.9	-2.4
8	2.4	-3.1	2.1	-2.8	2.3	-3.0
9	2.5	-3.6	2.6	-3.8	2.5	-3.6
Likelihood Ratio Chi-square	**1187**				**1209**	

Figure 5



From this result we conclude that the model predicts performance on Problem 1 for either group from that group's performance on the multiple choice anchor and the other group's performance on Problem 1. We view this as some support for the joint scaling that this exercise in model fitting has accomplished.

Discussion & Conclusions

> "Establish equating procedures with the highest level of precision practicable when scores on different test editions are intended to be comparable."
>
> *ETS Standards for Quality and Fairness*, 1987, p. 18

Why do we need to equate at all?

Current practice does not equate across items within a particular part of the exam. This means that if a student receives a raw score of 6 on a problem, it is of no consequence whether it was on problem 2 or 3. As we have shown, Problem 3 is more difficult, and so a student who gets a score of 6 on it is, in a very real sense, demonstrating more proficiency than another student who gets a 6 on Problem 2. Why is no adjustment made? One argument we have heard is that both students had the opportunity to answer either question, and if someone chose freely to attempt the harder problem it was their choice and they must live with the consequences of their action. We have some (but not a lot of) sympathy toward this view. The validity of this view hinges on the assumption that both students really had a choice. This may not be correct. The construction of the exam is meant to mirror the diversity of valid chemistry courses that might be offered. One course might emphasize stoichiometry, a second might emphasize thermodynamics. Both courses, and the viewpoints that generated them, may be equally valid. And yet, students who took the former course and confront the 'choice' really have no choice. The thermodynamics question is not one that they could successfully attempt.

Strenuous attempts are made in the course of test development to create problems of equal difficulty. We are certain that if empirical evidence were presented during the process of test development that a pair of problems differ significantly in their difficulty, modifications would be made to the problems to make them more equitable. Thus, we argue that when this evidence turns up after a test has been administered, canons of fairness and psychometric practice require some *post hoc* statistical adjustment. The procedures that we have developed and illustrated here are a reasonable place to begin such an adjustment.

Should we allow choice? If so, how many?

A broader question that needs serious consideration revolves around the use of a test format that allows examinee choice. As we pointed out earlier, this test has 1,120 different forms. The canon of good practice quoted at the beginning of this section

makes it clear that under such conditions these different forms must be equated as accurately as possible. Yet is any equating even possible? In this paper we have attempted to equate just two of these forms. We seem to have succeeded, but without the confirmatory evidence that can only be obtained with further data, we cannot be sure. The procedure we followed was time-consuming (in both human and machine time) but conceivably could be done for more than just two groups. The extent to which our results are convincing is partially due to the size of the sample used in the various estimations. Dividing the sample into 1,120 groups will surely leave some groups with vanishingly small samples. In those groups many of these analyses would lose their credibility.

In addition to difficulties associated with sample sizes, there is a substantial covariance between item difficulty parameters and the location of group proficiency distributions. This can lead to instabilities in the estimation process. In this study we found local extrema on the path to final parameter estimates. With just two groups a careful exploration of the multivariate response surface was possible and these potential snares avoided. In an analysis estimating hundreds of groups' means such scrutiny might not be possible. At least part of this problem is due to the way that MULTILOG estimates group means—as just another parameter in the likelihood equations. If this estimation was done separately we believe that these problems could be solved, although this is informed speculation at the moment and more work is required.

How good are the constructed response problems?

The Problems considered here appear to be very good indeed. We found in the analyses involving just two problems (1 & 2 or 1 & 3) that the reliability for each of these short tests was about .75. This is an indication of the amount of information that can be obtained with partial credit and a rigorous scoring scheme. Some of this high reliability can be chalked up to an easy task of separating a group of well trained examinees from another group of poorly trained ones (remember the abundance of zero scores). Nevertheless, these appear to be very good items. A lesson to be learned is that one can get a reliable test from constructed responses that are scored by human judges, if the rules that the judges use are specific and rigorous. This result should not be construed as support for holistic scoring. We do not know the extent to which these rigorous scoring rules can be automated, but to the extent they can be, then human judges are merely being used to read handwriting. Computerized administration of the test may solve the handwriting problem and automated scoring can ensue.

If the establishment of a rather rigidly defined set of scoring criteria is a *sine qua non* for reliable judging, how much do we gain using constructed responses over multiple choice items? The latter have many advantages in terms of time utilization, domain coverage, and practicality of use. Moreover, as we discovered, they seem to be testing very much the same thing. There are many contemporary arguments supporting a constructed response format. These usually involve the sorts of cognitive activity they require, as well as the structure of study that ensues if they are known to be in use.

These advantages may exist, but the empirical evidence observed from performance on these items shows very little deviation from a single factor.

It may be helpful to consider a broader view of the multiple choice item. A recent proposal (Johnson, 1991) suggested using ten choice items (to reduce the effects of guessing) in which the examinee's task is to choose the option closest to the correct answer (to prevent working backward). Thus a question "Find the square root of 5" might have as answers: 0, .5, 1, 1.5, 2, 2.5, 3, ... The extent to which this scheme is useful for a broader range of problems is unknown. We mention it only as an example of what can be achieved within the rubric of a completely objectively scored test format. Note that this scheme can incorporate expert judgement for more subtle kinds of questions (i.e., "Which statement best characterizes Molly in Joyce's *Ulysses*?"). Experts decide, subjectively, what are correct answers and then the question is graded objectively.

Meanwhile, to the extent that the use of constructed response problems is required, and to the extent that the responses may be characterized with a unidimensional IRT model, we have shown how comparable scores can be produced in the context of choice.

References

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261-280.

ETS Standards for quality and Fairness. (1987). Princeton, NJ: Educational Testing Service

Johnson, B. R. (1991). A new scheme for multiple-choice tests in lower-division mathematics. *American Mathematical Monthly, 98* (5), 427-429.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach,* Orlando, Florida: Harcourt Brace Jovanovich.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison Wesley.

Mann, H. (1845). *A description of a survey of the Grammar and Writing Schools of Boston in 1845*. Quoted in O. W. Caldwell & S. A. Courtis (1923). *Then and now in education,* Yonkers-on-Hudson, New York: World Book Company, p. 37-40.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement, 4,* Part 2, No. 17.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28*, 237-247.

Thissen, D. (1991). *Multilog user's guide* [computer program]. Chicago, IL: Scientific Software.

Thissen, D., Steinberg, L., & Fitzpatrick, A.R. (1989). Multiple-choice models: The distractors are also part of the item. *Journal of Educational Measurement, 26*, 161-176.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: A use of multiple-categ ical response models. *Journal of Educational Measurement, 26*, 247-260.

Thissen, D., Steinberg, L., & Wainer, H. (1992). Detection of Differential Item Functioning using the Parameters of Item Response Models. In P. W. Holland & H.

Wainer (Eds.) *Differential Item Functioning* . Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24,* 185-201.

Wainer, H., & Lewis, C. (1990). Toward a psychometrics for testlets. *Journal of Educational Measurement, 27,* 1-14.

Wainer, H., Sireci, S.G. & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement, 28,* 197-219.